CLAIMS

We claim:

1. A method of guaranteeing failure notification in a distributed system operating on a plurality of nodes in a network, the method comprising:

creating a failure notification group comprising the plurality of nodes, wherein the failure notification group has a unique identifier;

associating a failure handling method of an application with the unique identifier of the failure notification group;

ascertaining a failure; and

when the failure is ascertained, signaling a failure notification to each node in the failure notification group and executing the failure handling method.

2. The method of claim 1, further comprising disassociating the failure handling method from the unique identifier after the failure is ascertained and the failure handling method has been executed.

3. The method of claim 1, wherein creating a failure notification group includes:

verifying that each node in the failure notification group exists; and

generating the unique identifier for the failure notification group if each node in the failure notification group is successfully contacted.

4. The method of claim 3, wherein creating a failure notification group includes executing the failure handling method if each node in the failure notification group is not successfully contacted.

5. The method of claim 1, wherein creating a failure notification group includes:

generating the unique identifier for the failure notification group;

sending an invitation message containing an application state and the unique identifier to each node of the failure notification group; and

verifying that each member of the failure notification group received the invitation message.

6. The method of claim 5, further comprising, if any node in the group of nodes fails to receive the invitation,

signaling a failure notification to nodes that already received the invitation message; and

executing the failure handling method.

7. The method of claim 1, wherein signaling a failure notification includes sending a failure notification message to nodes in the failure notification group.

8. The method of claim 1, wherein signaling a failure notification includes failing to respond to a communication request from a node in the failure notification group.

9. The method of claim 1, wherein signaling a failure notification includes failing to respond only to communication requests related to a failure notification group for which a failure has been ascertained.

10. The method of claim 1, wherein ascertaining a failure includes ascertaining a failure in a communication link to at least one other node in the failure notification group.

11. The method of claim 1, wherein ascertaining a failure includes receiving from the application an instruction to signal the failure notification.

12. The method of claim 1, wherein ascertaining a failure includes having failed to repair the failure notification group one or more times.

13. The method of claim 1, wherein ascertaining a failure includes distinguishing between a communication failure between two nodes that are both in the failure notification group and a communication failure between two nodesthat are not both in the failure notification group.

14. The method of claim 1, wherein the failure is ascertained from an application pinging each node in the failure notification group, and determining the failure when a response to a ping is not received.

15. The method of claim 1, wherein the nodes in the failure notification group have a spanning tree topography, wherein the failure is ascertained from an application pinging adjacent nodes in the spanning tree, and determining the failure when a response to a ping is not received.

16. The method of claim 1, wherein the nodes in the failure notification group are a subset of nodes in an overlay network, wherein creating a failure notification group includes creating a multicast tree by sending a construction message to each node in the failure notification group.

17. The method of claim 16, wherein the construction message is routed to each node in the failure notification group through an overlay routing path, and nodes in the overlay routing path record pointers to adjacent nodes in the overlay routing path.

18. The method of claim 16, further comprising receiving a confirmation message, wherein the construction message is routed to each node in the failure notification group through an overlay routing path, and upon receiving the confirmation message, each node in the overlay routing path records a pointer a preceding node, and wherein the confirmation message is routed through the overlay routing path in reverse, and upon receiving the confirmation message, each node in the reverse overlay routing path records a pointer to a preceding node.

19. The method of claim 16, wherein ascertaining the failure includes ascertaining that a communication link to a node in the overlay network has failed, and determining whether the node was a member of the multicast tree.

20. The method of claim 19, wherein if the node was a member of the multicast tree, signaling a failure notification to adjacent nodes in the multicast tree.

21. The method of claim 19, wherein if the node was a member of the multicast tree, signaling a failure notification to adjacent nodes in the multicast tree by not responding to messages from the adjacent nodes.

22. The method of claim 19, wherein if the node was a member of the multicast tree, executing the failure handling method.

23. A method of guaranteeing failure notification in a distributed system operating on a plurality of nodes in a network, the method comprising:

receiving a unique identifier for a failure notification group, the failure notification group comprising the plurality of nodes;

associating a failure handling method of an application with the unique identifier of the failure notification group;

ascertaining a failure; and

when the failure is ascertained, signaling a failure notification to each node in the failure notification group and executing the failure handling method.

24. The method of claim 23, further comprising performing garbage collection to disassociate the failure handling method from the application state after the failure is ascertained and the failure handling method is executed.

25. The method of claim 23, wherein signaling a failure notification includes sending a failure notification message to nodes in the failure notification group.

26. The method of claim 23, wherein signaling a failure notification includes failing to respond to a communication request from a node in the failure notification group.

27. The method of claim 23, wherein signaling a failure notification includes failing to respond to only communication requests related to a failure notification group for which a failure has been ascertained.

28. The method of claim 23, wherein ascertaining a failure includes ascertaining a failure in a communication link to at least one other node in the failure notification group.

29. The method of claim 23, wherein ascertaining a failure includes receiving from the application an instruction to signal the failure notification.

30. The method of claim 23, wherein ascertaining a failure includes having failed to repair the failure notification group one or more times.

31. The method of claim 23, wherein the failure is ascertained from an application pinging each node in the failure notification group, and determining the failure when a response to a ping is not received.

32. The method of claim 23, wherein the nodes in the failure notification group have a spanning tree topology, wherein the failure is ascertained from an application pinging adjacent nodes in the spanning tree, and determining the failure when a response to a pin is not received.

33. The method of claim 23, wherein the nodes in the failure notification group are a subset of nodes in an overlay network, further comprising joining a failure notification tree, including:

receiving a construction message from a creator node through an overlay routing path; and

recording a pointer to adjacent nodes in the overlay routing path.

34. The method of claim 33, further comprising sending a confirmation message to the creator node, wherein the construction message is routed to each node in the failure notification group through an overlay routing path, and upon receiving the confirmation message, each node in the overlay routing path records a pointer a preceding node, and

wherein the confirmation message is routed through the overlay routing path in reverse, and upon receiving the confirmation message, each node in the reverse overlay routing path records a pointer to a preceding node.

35. The method of claim 33, wherein ascertaining a failure includes distinguishing between a communication failure between two nodes that are both in the failure notification group and a communication failure between two nodes that are not both in the failure notification group.

36. The method of claim 33, wherein ascertaining the failure includes ascertaining that a communication link to a node in the overlay network has failed, and determining whether the node was a member of the multicast tree.

37. The method of claim 36, wherein if the node was a member of the multicast tree, signaling a failure notification to adjacent nodes in the multicast tree.

38. The method of claim 36, wherein if the node was a member of the multicast tree, signaling a failure notification to adjacent nodes in the multicast tree by not responding to messages from the adjacent nodes.

39. The method of claim 36, wherein if the node was a member of the multicast tree, executing the failure handling method.

40. A method of guaranteeing failure notification in a distributed system operating on a plurality of nodes in a network, wherein the plurality of nodes are a subset of nodes in an overlay network, the method comprising:

joining a failure notification tree;

ascertaining a failure in a communication link to an adjacent node in the tree; and

signaling a failure notification when the failure is ascertained.

41. The method of claim 40, wherein joining the failure notification tree includes:

receiving a first message from a creator node through an overlay routing path;

recording a pointer to an overlay node from which the first message was received; and

forwarding the first message to a node in the failure communication group via a next node in the overlay routing path.

42. The method of claim 41, further comprising recording a pointer to the next node.

43. The method of claim 41, wherein joining the failure notification tree further includes:

receiving a second message from the node in the failure notification group through the overlay routing path;

recording a pointer to an overlay node from which the second message was received; and

forwarding the second message to the creator node via the overlay node from which the first message was received.

44. The method of claim 40, wherein ascertaining a failure includes distinguishing between a communication failure between two nodes that are both in the failure notification group and a communication failure between two nodes that are not both in the failure notification group.

45. The method of claim 40, wherein ascertaining a failure includes having failed to repair the failure notification group one or more times.

46. The method of claim 40, wherein ascertaining the failure includes ascertaining that a communication link to a node in the overlay network has failed, and determining whether the node was a member of the multicast tree.

47. The method of claim 46, wherein if the node was a member of the multicast tree, signaling a failure notification to adjacent nodes in the multicast tree by not responding to messages from the adjacent nodes.

48. An application program interface embodied on computer readable media, comprising:

a first function for creating a failure notification group and assigning a unique identifier to the failure notification group;

a second function for associating a failure handling method of an application to the unique identifier; and

a third function for signaling a failure notification to the failure notification group.

49. The application program interface of claim 48, the first function comprising a first parameter representing a set of nodes and a second parameter returning the unique identifier that is a result of the first function.

50. The application program interface of claim 48, the first function comprising a first parameter representing a set of nodes, a second parameter representing an application state, and a third parameter returning the unique identifier that is a result of the first function.

51. The application program interface of claim 48, the second function comprising a first parameter representing the failure handling method and a second parameter representing the unique identifier.

52. The application program interface of claim 48, the third function comprising a first parameter representing the unique identifier.